

## Topic: Home Based Trip Estimation from Mobile Phone Data

### Abstract

Traditionally, forecasts of travel demand have been made based on the data gathered from manually collected surveys. Recent advances in communication technologies enable the generation of Spatio -Temporal data and have proven to be a valuable source of data for human behavior studies. This work builds on the literature on transport modelling, aiming to evaluate the feasibility of CDR as raw data for travel demand forecasts. CDR Data was collected from mobile phone users in the Western Province of Sri Lanka over three months, and the study proposed a method to extract Home-Based Trips from collected CDR data. The method follows steps including noise reduction from CDR, significant location identification, trajectory extraction from labelled CDR data. The result showed that the correlation exists with traditionally collected travel data, it tends to increase at a higher granularity and for commute travel, such as Home-Based Work trips.

**Keywords:** CDR Data, Mobile data, Four step model, Trip generation, Load Sharing Effect

### 1 Introduction

Current practice, travel demand estimation begins with the collection of personal travel data through instruments such as home visit surveys where record of trips is made for each member of a household in terms such as origin, destination, purpose, travel mode, time of trip etc. Such data is thereafter processed a sequential four-step model (Ortúzar & Willumsen, 2011) , consisting of a Trip Generation, Trip Distribution, Mode choice and Route Assignment stages to estimate the travel demand and its patterns usually aggregated over a community.

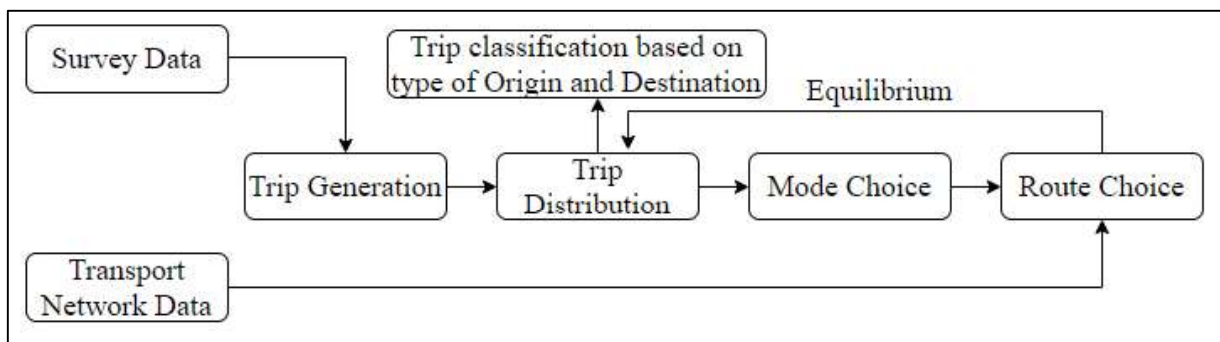


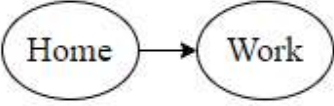
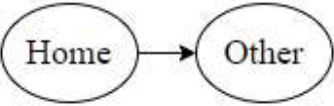
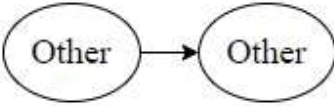
Figure 1: Four-Step Model

As in figure 1, the first stage of the classical transport model aims at predicting the total number of trips generated from a zone and attracted to each zone of the study area. In general, the no. of trips generated are calculated directly from the no. of trips made by the household who reside in each zone or with some other properties of the zone like population, employment, car ownership etc. Similarly trip attractions are affected by the no. of employment, schools and etc. Though productions and attractions give an idea of the level of trip making it is important to identify from where to where the trips take place. The second stage of four step model address this by storing trips made from an origin to a destination during a particular period in a trip matrix, commonly introduced as the Origin Destination matrix. The Origin Destination matrix can be disaggregated

by person type or purpose. Further in order to get a better idea on the trip making patterns, the third and fourth steps focus on the mode of transport chosen and route taken respectively. (Ortúzar & Willumsen, 2011)

The modelled trips are typically classified as Home-Based Work, Home-Based Other and Non-Home-Based trips as illustrated in Table 1.

Table 1: Definition of Home - Based Trips

| Home-Based Work  | Home-Based Other  | Non - Home-Based   |
|--|---|--|
|   |    |                                     |
| Trips that have one end at home and the other end at a place of work. These are usually regular and have repetitive and cyclical patterns. | Trips that have one end at home and the other end at a non-home location other than work. The regularity of the trips are much less compared to Home Based Work trips | Trips that have both ends at non-home locations. These trips are much irregular and does not depict a cyclical patten. |

Since surveys required for four-stage modelling are expensive to conduct and laborious, they are usually infrequent (Stopher & Greaves, 2007) resulting in large time gaps between surveys making them outdated even though the data at point of collection is rich in mobility information. Besides, the collected data can be misguided since the survey participants must recall information on past travel when filling the questionnaire.

The availability of digitally generated behavioral data pertaining to mobility of a community will enable the estimation of travel demand at more frequent intervals as well as at lower cost and at greater accuracy. Mobile phone data also known as Call Detail Records (CDRs) has the potential to provide such information that can fundamentally change, how planners analyze the demand for mobility and for existing transport systems. CDRs have been previously used widely in studies to explore the knowledge of individual human mobility (Ahas et al., 2010; Järv et al., 2014; Jiang et al., 2016; Leng, 2016)

CDRs are generally collected by mobile operators for internal billing purposes on a large scale and stored for an extended period. This enables the convenient and fast deployment of the data for travel estimation. CDR have a number of advantages that improve the accuracy over home visit data namely being, (a) more representative than sampled traditional data since everyone with a smart phone generates data; (b) able to capture more variability over a longer period of time whereas home visit data will be for a day or a week at most and (c) not reliant of human memory from which home visit surveys are compiled..

Despite the advantages, there are significant challenges when deriving mobility patterns using CDRs. There are specific difficulties in data storage, data sharing and removing the noise during the analysis that will be discussed in this paper. Moreover, CDRs cannot be used by the traditional four step models and required a different method altogether.

Primary objective of this paper is to establish a methodology of using Call Detail Records in understanding individual mobility records. This paper captures the first two steps in the four-step travel demand estimation model by firstly estimating both trip generation and attractions from Home-Based Trips and secondly by estimating the OD matrix of travel in a given area. Generally, trip generation models establish a mathematical relationship between trip making rates and the demographics of individuals or households in a given area (Ortúzar & Willumsen, 2011). Since the CDRs are obtained as individual data, the trips patterns are aggregated over an area such as an administrative district or a Divisional Secretariat (DS) Division. Furthermore, each of the trips observed to have the home location as one of the ends of the trip further categorized as (a) Home-based Work, (b) Home-based Other and (c) Non-Home-Based Trips. Broadly, the paper discusses how mobile phone data can be used for developing an OD matrix.

## **2 Literature Review**

CDRs being intended for recoding the tower which picks up the phone call, the caller number, time of making the call and its duration, they are not as structured to capture salient mobility data that can meaningfully be directly used to determine if a trip has been made and where its origin and destinations were, what time the trip was made or the route through which the trip was made (Ravulaparthi et al., 2016). The compilation of an Origin-Destination (OD) Matrix is an important output of such data. Although CDRs are often sparse in space and time, the large volume and longer observation period of mobile phone data can infer human footprints on an unprecedented scale (Lim et al., 2013). CDRs can reasonably represent spatiotemporal information of mobile phone users' movements at cellular-tower or much finer-grained level, depending on the location positioning technology employed by service carriers (Jiang et al., 2016). The CDRs therefore possess the potential to estimated mobility information ranging from trip generation identification to OD matrix estimation to mobility pattern recognition. This section of the paper reviews previous studies that use CDR to derive human mobility.

Iqbal et al. (Yang et al., 2017) use CDRs collected in Dhaka, Bangladesh over a period of one month, combined with traffic counts data, to estimate tower-based transient origin-destination (OD) matrices for different periods and converts them to node-to-node transient ODs. Wang et al. (Jiang et al., 2016) follows a similar approach in developing ODs from data collected in Boston and San Francisco areas. Fekih et al. develop a method by which trips were extracted from the spatiotemporal traces of users based on a minimum stationary time assumption that enables to determine activity (stop) zones for each user (Fekih et al., 2020). Mamei et al, identified individual trips based on movements between cells (sequence of CDR) or of inferred routine trips known to be taking place. Then, individual trips are aggregated to create the OD matrix (Mamei et al., 2019). In a study conducted in Mumbai, CDR data are collected by day type and combined to generate mobility traces for each user by superimposing all activities' locations by each specific type of day.

The records are then aggregated for all users, multiplied by a scaling factor and converted to vehicle trips to arrive at an O-D matrix (Bera & Rao, 2011). Apart from that, several studies have developed algorithms to assign caller locations to traffic nodes from mobile phone towers from which the calls have been made. Bwambale et al., predicts individuals' demographic group membership relativeness based on their phone usage characteristics and then uses this relativeness as weights inside a latent class model for trip generation, with different classes representing different socio-demographic groups. (Bwambale et al., 2017). These methods have shown the feasibility of using CDR data to develop trip ODs. However, these processes have considered the locations of two successive calls as segments of a full trip and have developed OD matrices between locations from which calls are made and not necessarily between physical trip origins and destinations. Though when aggregating some accuracy is achieved, due to cancellation of over and underestimation of trips, there is a flaw in the methodology.

Human's organize their daily travel based on activities and anchor locations that are important and add value to their daily life (Jiang et al., 2016). Most studies in use of CDRs consider the number of days they appear in a location and the number of calls made through a particular cell tower as inputs to derive an individual's meaningful or anchor locations (Ahas et al., 2010)(Kung et al., 2014). Some studies use clustering techniques, where spatially close cell towers are represented by clusters containing towers that frequently channel call activities (F. Wang & Chen, 2018). Mamei et al., Cluster CDRs together based on the spatial parameter and weigh clusters based on time to determine those associated with frequented places (Mamei et al., 2016). Luo et al, combine CDR data et al.road network data to identify trajectories aiming at decreasing base station positioning error. Next, the trajectory regularity is measured and home, work locations were estimated using the regularity (Luo et al., 2020) (Zagatti et al., 2018). Total travel can be, measured by aggregating user locations' preferences (Leng, 2016) and establishing the variance of such measures through statistical models (Järv et al., 2014).

Though many researchers have addressed this trip detection based on anchor points, the methods they have used do not capture the individualistic behaviors, where same regularity, frequency criteria are used for all the individuals in the sample or clusters. Combining these data with person-level attributes is more challenging, and previous studies have not developed data fusion frameworks to address this challenge.

Motivated by the above literature, this study uses a random sample of CDRs from the Western Province of Sri Lanka as a case study to develop a data processing opportunity. The study leads to extract individual mobility patterns to understand travel behaviour at the disaggregated level.

### **3 Study Area and Data**

This paper is based on the comparison of the trips ODs obtained by using the CDRs and the ODs obtained from the traditional manually complied Household Survey Data.

#### **3.1 CDR data**

The Call Detail Records were provided by one of the leading mobile operators in Sri Lanka through LIRNEasia - a regional ICT policy and regulation think-tank. CDRs were entirely

pseudonymized by the operator and identified with a security ID, so that the privacy of the users are not violated. This event-driven data is typically generated through incoming and outgoing voice calls and text messages and contains:

- A random ID number to identify a specific phone,
- Exact time and date of call or text message
- If a call, its duration
- Tower the call connected to when call was placed.

The cell tower that connects a call or the serving cell tower, is generally the tower nearest to the user. This enables the identification of proximity of a mobile phone user's location whenever a call is placed or received. In other words, the location of a user is defined as the cell tower connected, since that is the observable finest granularity. The understanding of the movements of individuals in the Western Province of Sri Lanka was pursued with a the CDRs of a random sample of 10,000 users in the database that was made available from CDRs for the three month period of May, June and July in 2013.

### Zones of Aggregation

There are a total of 763 cell towers in the dataset. Each cell tower has a unique ID with the corresponding latitudinal and longitudinal readings of the location.

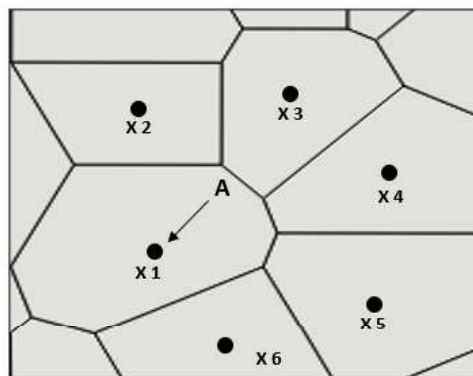


Figure 2: Example of a Voronoi Cell Distribution

As shown in Figure 2, the precision of the mobile device user's spatial positioning corresponds to the coverage area of a network antenna (Khan et al., 2015). The user can be anywhere within the coverage area, not spatially fixed but varies according to population density and socio-economic features.

## 3.2 Household Visit Surveys (HVS) data

The study uses the household visit survey (HVS) data to validate its findings. The HVS was conducted by the Ministry of Transport with the technical support of Japan International Cooperation Agency (JICA) as one of the largest and most comprehensive transport surveys carried out in Sri Lanka. The study area of the HVS covers the entire Western Province, which

includes the three administrative districts of Colombo, Kalutara and Gampaha consisting of 40 Divisional Secretariat Divisions (DSD) wherein there are 2,496 Grama Niladhari Divisions (GNDs). There was a total of 35,000 sample households randomly selected from the list of postal addresses obtained from the electoral registration list. The study ensured a sampling size of 4% across the province.

## 4 Methodology

This research proposes a methodology with multiple layers for processing the CDRs as illustrated in Figure 3 below. The methodology for converting the CDRs is achieved through six steps as explained below

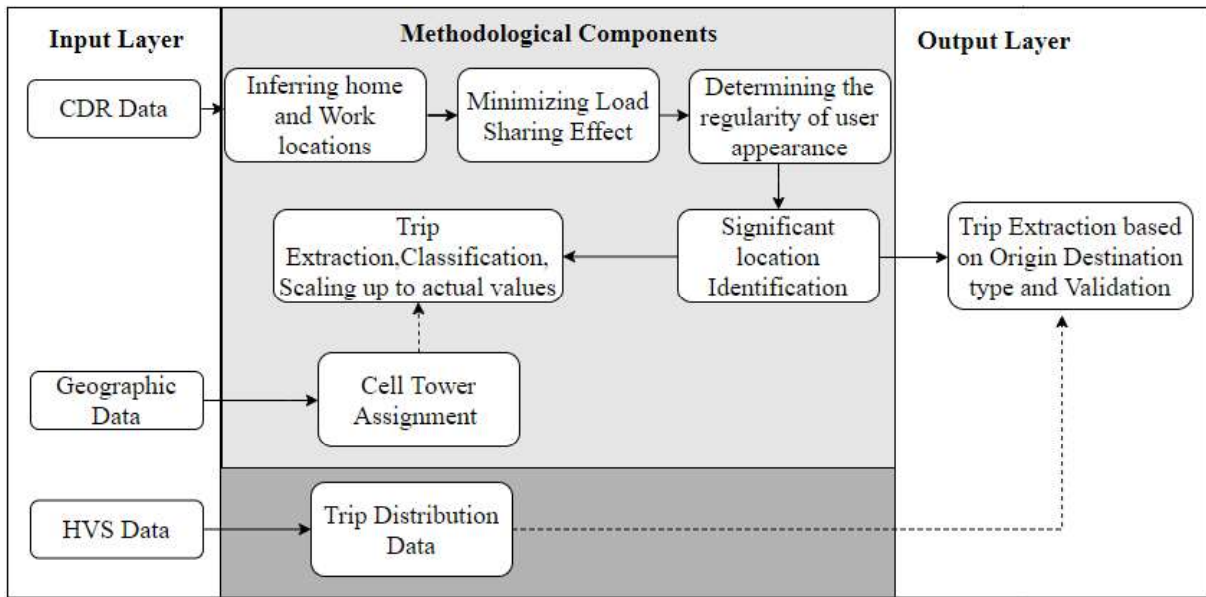


Figure 3: Research Methodology

### 4.1 Step 1: Inferring home and Work locations

Since the study aims to estimate home-based trips using CDRs, identifying a caller's home and other locations is of foremost importance. The methodology adopted in (Ahas et al., 2010), (Kung et al., 2014) has been used where Home has been identified as the location most frequented and wherein a user has relatively longer stays mostly during the nighttime and in most cases stays during weekends. Accordingly, the cell tower location that was used most frequently during the hours (8:00 p.m.–4:00 a.m.) was considered as the home location. Work locations were considered as a regular location frequented by a user during the daytime on weekday but generally not visited during weekends. In general, these were assumed to be locations where the user had a more extended stay duration or spent large blocks of time. This assumes that the workplace location to be in the cell tower area that was most frequently used during the core office hours (10:00 a.m.–4:00 p.m.) on weekdays. However, this does not enable a person who works close to home but in a different location to have the workplace identified separately from home.

## 4.2 Initial data processing

Initial data processing involves filtering out the noise in the CDRs generated due to phenomenon of load balancing function performed by the mobile operator. This occurs mainly because the operator diverts call traffic to adjacent towers handling lower call volumes at that movement to balance the load on its tower network. Due to this process, the call is assigned to a different tower other than the one to which the caller's location belongs. This can cause two subsequent locations of a particular caller to be indicated by a presence in two different but adjacent tower areas even though the caller had not moved physically. The study uses a speed - based filtering technique shown in Figure 4 to minimize the error created due to load sharing effect.

Accordingly, the CDRs are arranged chronologically for each user, where the callers speed of travel between two consecutive records is calculated. Calculating the speed requires the distance between the two tower locations and the time taken. Even though a user can be anywhere within the cell tower coverage area, it was assumed that the centroid of the cell tower connected as the user's location. Therefore, the available latitudinal and longitudinal data of the centroids are used to calculate the distance between the locations of two consecutive records embedded in the CDR data. Once the speed is calculated, any record with an irrational high travel speed is identified as a potential load shared record.

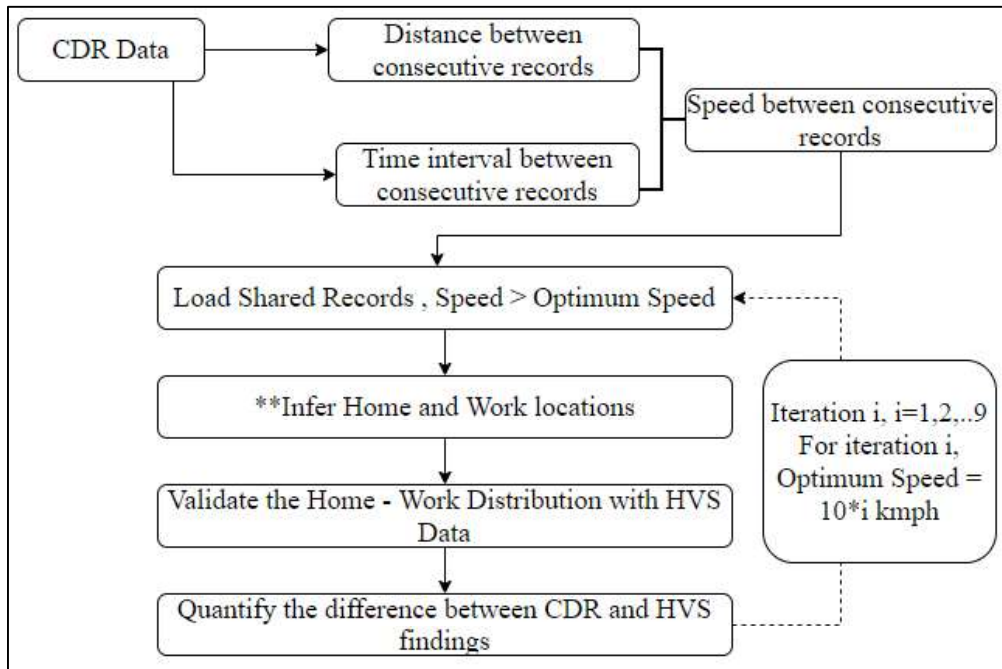


Figure 4: Methodology for Load Share Identification

To decide the threshold speed to filter the load shared records the users with clearly identifiable home and work locations are extracted, and their home DSDs and work DSDs are identified.. Since the speed between two consecutive call records is calculated, potential load shared records are removed for different speed limits varying from 10kmph-90kmph at a frequency of 10kmph. Respective work and home locations are identified at each speed, and the findings are compared with HVS data at each speed. Comparison was done at DSD level.



Findings are compared as percentages, Kullback–Leibler divergence which is a measure of how one probability distribution is different from a second is used to measure the difference in CDR and HVS distributions in each speed level. The speed limit which minimizes the error percentage is considered as the optimum speed to remove the load sharing effect in the CDR data.

As of the Kullback–Leibler divergence method, For discrete probability distributions P and Q defined on the same probability space, R, the relative difference from Q to P is defined to be,

$$D_{KL}(P||Q) = \sum_{x \in R} P(x) \log(p(x)/Q(x))$$

For the selected sample, the error was minimal at 40kmph and records with consecutive speeds higher than 40kmph are identified as load shared records, which accounted for 11% from the total records.

### 4.3 Identifying Significant locations

Individuals tend to make frequent visits to specific locations, while some places are visited infrequently. It is crucial to identify these significant locations or regularly visited locations which results in a higher number of calls from these locations more than other less-visited locations. This behavior of individuals divides the connected cell towers as regularly visited cell towers and irregular ones. The boundary of regularity and irregularity varies from one person to another, demonstrating that the cell regularity measuring to be individualistic. Once the regularly visited cells are identified; they can be labelled as significant or meaningful locations of the individual's behavior space as home or other.

#### 4.3.1 Measuring the regularity of cells

The study measures the regularity of visits made to each cell in terms of three parameters: (a) Regularity by time of the day (b), Regularity by day of the week and (c), regularity of the day of the month. For this purpose a single day was divided into different time segments ( $T_i$ ,  $i=1,2,...,10$ ) ( $T_1= 6am-8am$ ,  $T_2= 8am-10am$ ,  $T_3= 10am-12pm$ ,  $T_4= 12pm-2pm$ ,  $T_5= 2pm-4pm$ ,  $T_6= 4pm-6pm$ ,  $T_7= 6pm-8pm$ ,  $T_8= 8pm-10pm$ ,  $T_9= 10pm-4am$ ,  $T_{10}= 4am-6am$ ). The regularity of a particular user (i) being in cell (x) during the time category  $T_i$  is directly proportional to the three defined parameters a, b, and c. Table 2 shows an example of a user.

“a” = Total no. of days the user n has visited cell tower x during time category  $T_i$ .

“b” = Total no. of days the user n has visited cell tower x at time category  $T_i$  per week.

“c” = Total no. of days the user n has visited cell tower x at time category  $T_i$  per month.

The regularity of a particular user (n) at cell (x) in time category  $T_i \propto a*b*c$

Table 2: Measuring the regularity of cell towers

| User ID   | Cell_ID | Time category | No. of days (a) | Total no. of days per week (b) | Total no. of days per month (c) | Regularity Measure ( $a*b*c$ ) |
|-----------|---------|---------------|-----------------|--------------------------------|---------------------------------|--------------------------------|
| 685716430 | 26889   | 6pm-8pm       | 59              | 4                              | 19                              | 4484                           |
| 685716430 | 26889   | 8pm-10pm      | 55              | 3                              | 18                              | 2970                           |



|           |       |           |    |   |    |      |
|-----------|-------|-----------|----|---|----|------|
| 685716430 | 26889 | 10pm-4am  | 48 | 3 | 16 | 2304 |
| 685716430 | 26889 | 4pm-6pm   | 27 | 1 | 9  | 243  |
| 685716430 | 26889 | 12pm-2pm  | 25 | 1 | 8  | 200  |
| 685716430 | 26889 | 2pm-4pm   | 24 | 1 | 8  | 192  |
| 685716430 | 26889 | 10am-12pm | 21 | 1 | 8  | 168  |

Once the regularities of visiting each cell for each time category is calculated the regularity measure with the highest score was used to extract the significant locations from the available list of all possible cells.

#### 4.3.2 Extracting the Significant Locations

As explained earlier, cell regularity measuring method should be individualistic. Mathematical Linkage Analysis (MLA) derived from graph theory can be used as a technique to address each user's behavior separately (Järv et al., 2014). Initially, cells of each user are arranged in descending order based on the regularity of cells which is introduced as the real configuration. MLA compares the real configurations to idea-typical configuration in which the regularity is similar over exactly first, second, third... cells. In other words, in the first ideal-typical condition only one cell is there; in the second situation two cells are there where the regularity is equal in both cells. Next, the R-sq between the real distribution and each of the ideal-typical condition is calculated. The no of cell in the iteration with the largest R-sq value gives the number of significant locations of the individual. Table 3 indicates a user's example, where the maximum R-sq value reaches in the second iteration. Therefore the no. of significant locations can be concluded as two for this considered user. Same process was carried out for the sample and identified the significant locations of each user.

*Table 3:Identifying Significant Locations*

| <b>Actual distribution of Regularity</b> |                                   |                                | <b>Ideal-Typical Configuration</b> |             |             |             |             |
|--|-----------------------------------|--------------------------------|------------------------------------|-------------|-------------|-------------|-------------|
| <i>Cell ID</i>                           | <i>Maximum Regularity measure</i> | <i>% of Regularity Measure</i> | <i>1(%)</i>                        | <i>2(%)</i> | <i>3(%)</i> | <i>4(%)</i> | <i>5(%)</i> |
| A  | 4880                              | 34%                            | 100%                               | 50%         | 33.33%      | 25%         | 20%         |
| B  | 3444                              | 24%                            | 0%                                 | 50%         | 33.33%      | 25%         | 20%         |
| C  | 574                               | 4%                             | 0%                                 | 0%          | 33.33%      | 25%         | 20%         |
| D  | 574                               | 4%                             | 0%                                 | 0%          | 0%          | 25%         | 20%         |
| E  | 573                               | 4%                             | 0%                                 | 0%          | 0%          | 0%          | 20%         |
| F  | 430                               | 3%                             | 0%                                 | 0%          | 0%          | 0%          | 0%          |
| G  | 429                               | 3%                             | 0%                                 | 0%          | 0%          | 0%          | 0%          |
|  |                                   | Rs-q                           | 0.6563                             | 0.9410      | 0.5259      | 0.306       | 0.1676      |

#### 4.4 Trip Extraction and Classification

Extracting physical movements of users or trip identification is a key task in this analysis and a specially focused objective since the trips are inferred based on a set of CDRs which are sparse in time and space. The process of turning consecutive sparse call records is referred to as the trip extraction for which a trip is defined as a movement between spatially distinct locations identified

as shown earlier namely home, work or significant location. Table 4 gives an illustration of identifying a trip.

Furthermore, the identified trips are classified into three segments based on the type of origin and destination as introduced in table 1. Such that when one end of the trip is home and the other is work, that movement was considered as a Home-Based Work trip. Trips where one end is home and other is a significant location other than work was classified as Home Based Other and trips where either end are non-home were grouped as Non-Home Based trips.

Table 4: Trip Extraction from Call Records

| Cell # | Cell Type Identified as  | Time of Appearance |   |
|--------|--------------------------|--------------------|---|
| A      | Work                     | 01/05/2013- 15:30  | } |
| A      | Work                     | 01/05/2013- 16:40  |   |
| B      | Home                     | 01/05/2013- 18:10  | } |
| C      | Significant location     | 01/05/2013- 19:30  |   |
| D      | Significant location     | 01/05/2013- 20:15  | } |
| E      | Non-Significant location | 01/05/2013- 22:00  |   |

A Trip (**HBW**)

A Trip (**HBO**)

Not a Trip

#### 4.5 Estimation of Origin-Destination Matrices

Travel demand analysis requires an understanding of aggregate trip patterns between different travel zones. In order to determine this each user's trips are aggregated to understand the cumulative travel between different travel zones over the time period for which the CDRs are available. This is enabled by assigning the cell towers to such travel zones covering the geographic area for which the CDRs are available. The granularity of the household survey data available for validation enables identification of travel zones at a (a) administrative district and (b) Divisional Secretariat Division (DSD) level. Therefore, a layer of cell towers represented by the respective latitudinal, longitudinal data are applied to a layer of the district and DSD boundaries using GIS software so as to superimpose the cell towers within the transport zones as shown in Figure 5

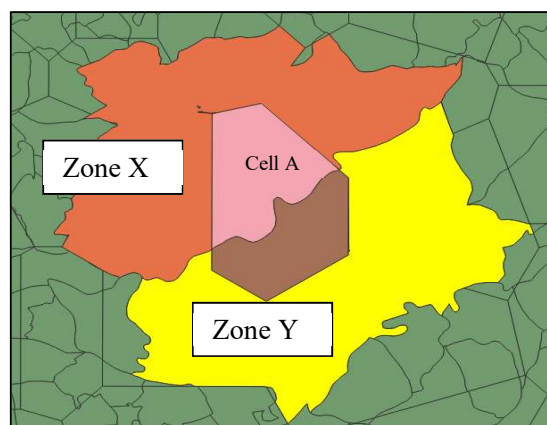


Figure 5: Assigning Cells to geographical areas

Since the cell tower boundaries are not coincident with those of the travel zones some apportionment of areas have to be made. Therefore, trip generations and attractions from a cell tower are proportionated based on the overlapping areas of boundaries. Then the trips will be calculated as follows,

Total Area of cell A = Area covered in Zone X ( $A_{Zone X}$ ) + Area covered in Zone Y ( $A_{Zone Y}$ )

Assume that cell A generates T no. of trips,

Total no. of trips generated from A that initiated from Zone X =  $T * [A_{Zone X} / (A_{Zone X} + A_{Zone Y})]$

Total no. of trips generated from A that initiated from Zone Y =  $T * [A_{Zone Y} / (A_{Zone X} + A_{Zone Y})]$

Cell towers are assigned based on the above method and defined the origins, destinations as Districts and DSDs. When the origins and destination of trips are aggregated by the travel zones they are also identified by their trip type namely as HBW, HBO or NHB. Consequently, daily OD matrices were developed for each trip type and each segment is compared at intra, inter District level and inter DSD level. Such that total trips made inside the district and within districts are compared.

Furthermore, Home Based Work trips which are regular in nature are expanded to actual trip counts by proportionating with the population of the Western Province. Table 5 shows an example of the expansion process for a single cell in the district level OD Matrix. Trip counts were expanded at district level (Both Inter and Intra) and DSD level (Inter).

Table 5: Expanding to Actual Trip Counts

| HVS Data   |    | CDR Data   |                    |
|--|----|--|--------------------|
| Total Population in Western Province:            | P1 | Total Population in Western Province as of the CDR sample:   | P2                 |
| No. of Home - based Work trips from zone X to Y: | C1 | No. of Home - based Work trips from zone X to Y for the considered period (Counting Weekdays only) | C2                 |
|  |    | No. of Weekdays in the considered period   | D                  |
|  |    | No. of Home - based Work trips per day from zone X to Y for the expanded population**              | $(C2/(P2*D)) * P1$ |

\*\* Assumption: Every individual who makes a work trip is assumed to have a mobile phone.

## 5 Results of CDR Estimation and Validation Using HVS data

Once the trips are extracted and classified as explained in the methodology, it could be observed that the trip proportions generated from CDR data are comparable with the HVS survey data. It is important to note that the trips are identified by assigning cells based on proportionating the cell areas to relevant geographic zones as explained the methodology, since it generates more accurate results at disaggregate level, rather than assigning cells alternatively to geographical areas.

*Table 5: Comparison of Trip Proportions*

| Trip Type        | Proportion of trips                  |   |
|------------------|--------------------------------------|---|
|                  | CDR Data (Per Day for 10,000 sample) | HVS Data (Per Day for Western Province) |
| Home Based Work  | 422 (26%)                            | 2700447 (27%)                           |
| Home Based Other | 813 (51%)                            | 5160050 (52%)                           |
| Non-Home Based   | 365 (23%)                            | 2126235 (21%)                           |

The home - based work trips which are much regular in nature are expanded to actual trip counts as explained in the methodology and table 6 shows the results generated at district level. R-Sq value of 97% could be observed at district level and standard error was 19%. Generated R-Sq values in trip generations and attractions at DSD level were 80%, 85% and accounted for a standard error of 38%, 33% respectively (Figure 5). It could be observed that the model fit decreases at disaggregate level due to the errors in allocating tower zones to travel zones which increases when the size of the travel zone decreases.

*Table 6: HBW Trips scaled up to actual population.*

| OD Pair (District level) | CDR Data expanded | HVS Data |
|--------------------------|-------------------|----------|
| Colombo-Colombo          | 290039            | 1150559  |
| Colombo-Gampaha          | 18658             | 152107   |
| Colombo-Kalutara         | 10725             | 71792    |
| Gampaha-Colombo          | 21538             | 154907   |
| Gampaha-Gampaha          | 148854            | 763714   |
| Gampaha-Kalutara         | 206               | 3388     |
| Kalutara-Colombo         | 12076             | 72925    |
| Kalutara-Gampaha         | 73                | 3466     |
| Kalutara-Kalutara        | 68374             | 327588   |

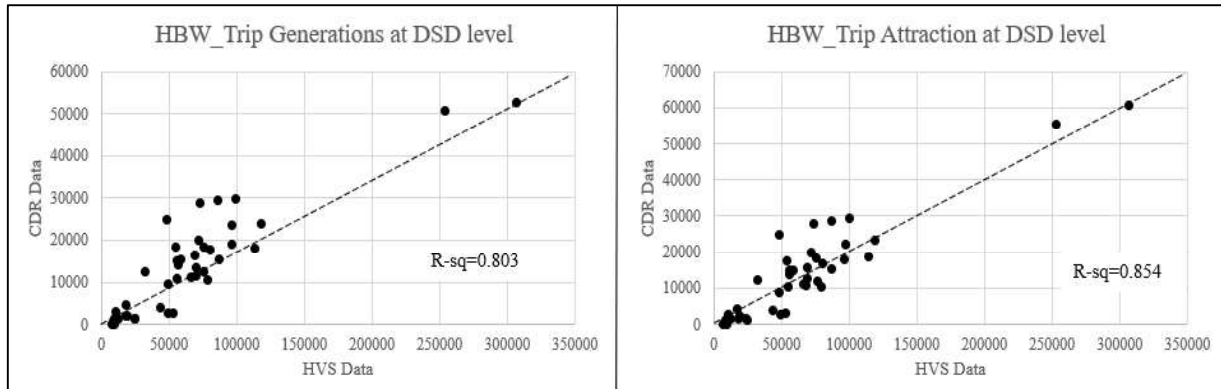


Figure 5: HBW trip validation at DSD level

Figure 6 shows the proportional validation of the other two trip types; HBO and NHB trips at inter and intra district level. Figure 7 and Figure 8 presents the comparison of trip generations and attractions at DSD level respectively. Both trip types indicate a reasonable R-Sq value at two geographical level and as of the HBW trips the accuracy seems to decrease at disaggregate level. But Home - based Work trips showcase the highest match. Higher regularity of home - based work trips can be considered as the possible reason for this similarity.

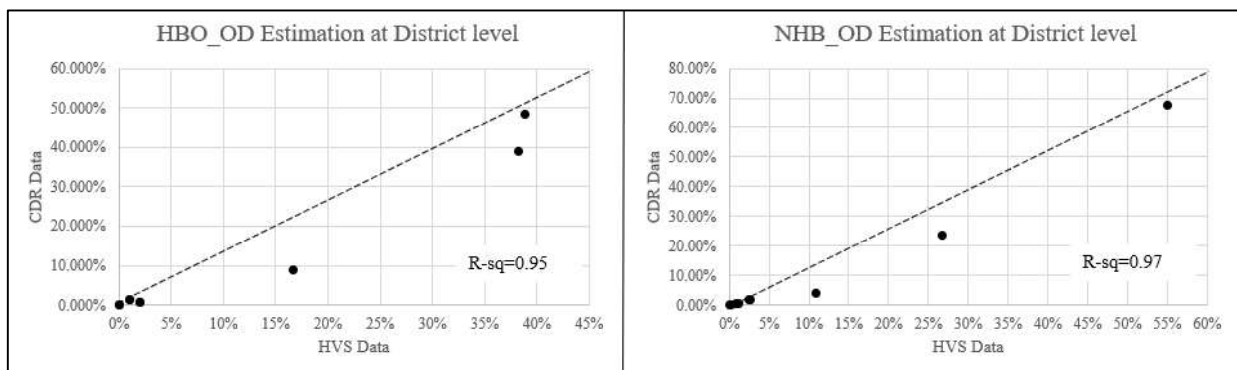


Figure 6: Validation at District level

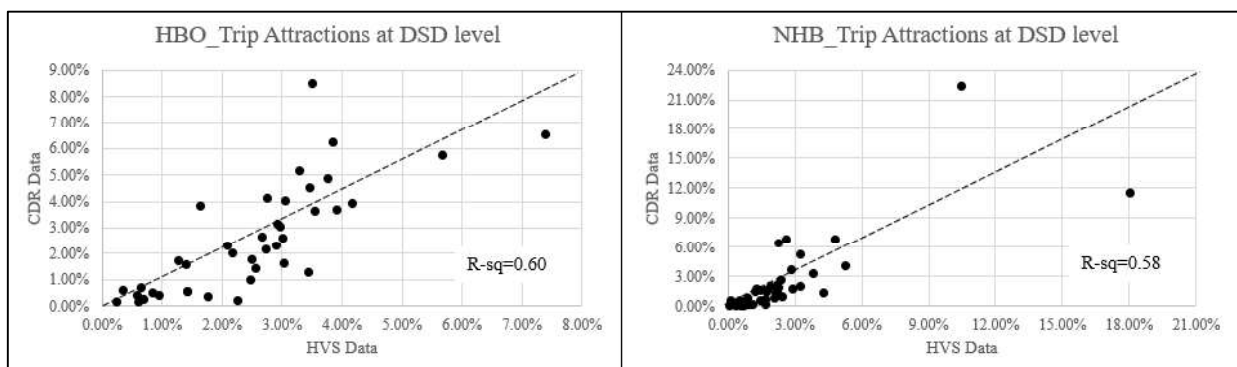


Figure 7: Trip attraction validation at DSD level

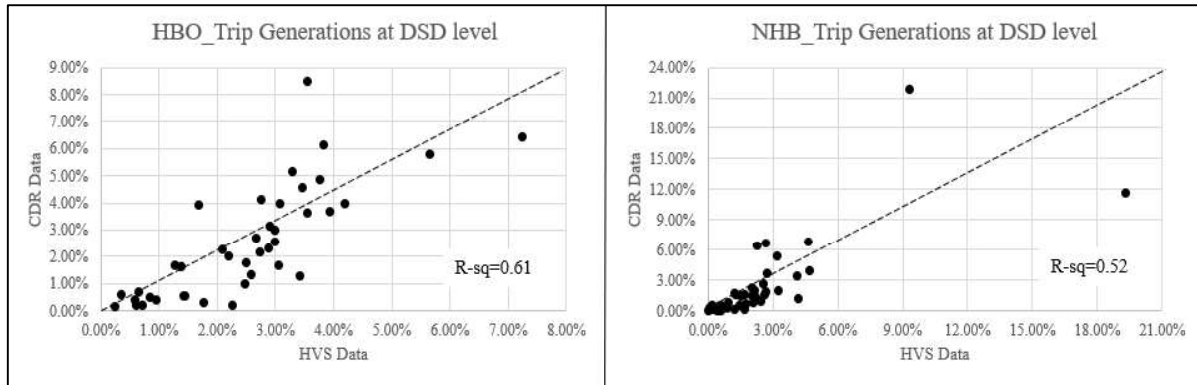


Figure 8: Trip Generation validation at DSD level

## 6 Conclusion

This paper describes a six-step methodological approach to formulate a home-based trip Origin-Destination (OD) matrix from Call Detail Records (CDRs). The technique when applied to CDRs of 10,000 users in the Western Province of Sri Lanka and validated with the daily trips estimated from Home Visit Survey (HVS) data collected for the same period gives acceptable results. This study reveals four important interventions required for this proceed to succeed. First, there is a need to separate load shared records using an optimum speed threshold so that the error created in tower switching operations can be minimized. Second, it is necessary to identify a users' significant locations using multiple linkage analysis to identify the regularity of an individuals' appearances in specific tower locations. Third, the study defines a method to classify a trip type as home-based work, home-based other and non-home based after identifying the movements between locations identified as home and work based on the regularity and timing of their appearances at such locations. Fourth, it concluded that the spatial granularity used to aggregate trips is an essential factor in the accuracy.

The paper concludes that the above method can be used successfully to determine Origin-destination matrices by trip type from CDRs thus enabling more accurate and regular estimation of travel at a fraction of the cost compared to traditional methods of deriving ODs from home-visit surveys. CDR analysis can be extended further by applying the methodology to Spatio-temporal data of different time frames and obtaining route choice, mode-specific trip tables, and significantly to identify public transportation.

## Acknowledgment

This research was supported by the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education funded by the World Bank.

## 7 References

- Ahas, R., Silin, S., Järvi, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17(1), 3–27. <https://doi.org/10.1080/10630731003597306>
- Bwambale, A., Choudhury, C. F., & Hess, S. (2017). Modelling trip generation using mobile phone data : A latent demographics approach. *Journal of Transport Geography*, June, 1–11.

<https://doi.org/10.1016/j.jtrangeo.2017.08.020>

- Fekih, M., Bellemans, T., Smoreda, Z., & Bonnel, P. (2020). A data - driven approach for origin – destination matrix construction from cellular network signalling data : a case study of Lyon region ( France ). In *Transportation* (Issue 0123456789). Springer US.  
<https://doi.org/10.1007/s11116-020-10108-w>
- Iqbal, S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin – destination matrices using mobile phone call data. *TRANSPORTATION RESEARCH PART C*, 40, 63–74. <https://doi.org/10.1016/j.trc.2014.01.002>
- Järv, O., Ahas, R., & Witlox, F. (2014). Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38, 122–135.  
<https://doi.org/10.1016/j.trc.2013.11.003>
- Jiang, S., Ferreira, J., & Gonzalez, M. C. (2016). Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. *IEEE Transactions on Big Data*, 3(2), 208–219. <https://doi.org/10.1109/tbdata.2016.2631141>
- Khan, F. H., Ali, M. E., & Dev, H. (2015). A hierarchical approach for identifying user activity patterns from mobile phone call detail records. *Proceedings of 2015 International Conference on Networking Systems and Security, NSysS 2015*.  
<https://doi.org/10.1109/NSysS.2015.7043535>
- Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS ONE*, 9(6).  
<https://doi.org/10.1371/journal.pone.0096180>
- Leng, Y. (2016). *Urban computing using call detail records : mobility pattern mining, next-location prediction and location recommendation*.  
<https://dspace.mit.edu/handle/1721.1/104156#files-area>
- Lim, J. S. E., King, I., Yu, P. S., & Nejd, W. (2013). *Behavior and* (Issue August).
- Luo, X., Zhou, Y., Yang, Y., & Wu, S. (2020). Research on Home and Work Locations Based on Mobile Phone Data. *Journal of Physics: Conference Series*, 1486(5).  
<https://doi.org/10.1088/1742-6596/1486/5/052013>
- Mamei, M., Bicocchi, N., Lippi, M., Mariani, S., & Zambonelli, F. (2019). Evaluating origin–destination matrices obtained from CDR data. *Sensors (Switzerland)*, 19(20), 1–17.  
<https://doi.org/10.3390/s19204470>
- Mamei, M., Colonna, M., & Galassi, M. (2016). Automatic identification of relevant places from cellular network data. *Pervasive and Mobile Computing*, 31(April 2014), 147–158.  
<https://doi.org/10.1016/j.pmcj.2016.01.009>
- Ortúzar, J. de D., & Willumsen, L. G. (2011). Modelling Transport. In *Modelling Transport*.  
<https://doi.org/10.1002/9781119993308>
- Ravulaparthi, S. K., Konduri, K. C., & Goulias, K. G. (2016). Fundamental linkages between activity time use and subjective well-being for the elderly population: Joint exploratory



- analysis framework for in-home and out-of-home activities. *Transportation Research Record*, 2566, 31–40. <https://doi.org/10.3141/2566-04>
- Stopher, P. R., & Greaves, S. P. (2007). Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice*, 41(5), 367–381. <https://doi.org/10.1016/j.tra.2006.09.005>
- Wang, F., & Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*, 87(December 2017), 58–74. <https://doi.org/10.1016/j.trc.2017.12.003>
- Wang, M., Schrock, S. D., Vander, N., & Mulinazzi, T. (2013). *Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data*. 76–86. <https://doi.org/10.1007/s13177-013-0058-8>
- Zagatti, G. A., Gonzalez, M., Avner, P., Lozano-Gracia, N., Brooks, C. J., Albert, M., Gray, J., Antos, S. E., Burci, P., zu Erbach-Schoenberg, E., Tatem, A. J., Wetter, E., & Bengtsson, L. (2018). A trip to work: Estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR. *Development Engineering*, 3, 133–165. <https://doi.org/10.1016/j.deveng.2018.03.002>